A Work Project, presented as part of the requirements for the Award of a Masters Degree in Economics from the NOVA – School of Business and Economics

UNDER THE VEIL OF CITATIONS

Diogo José Fernandes de Carvalho #443

Direct Research supervised by:

Professor Pedro Portugal

6th January 2012

UNDER THE VEIL OF CITATIONS

Diogo J. F. de Carvalho*

Supervised by: Prof. Pedro Portugal

ABSTRACT

Academics are often ranked on citation counts', which is considered an adequate proxy for author's quality and reputation. This paper seeks to find what is behind a cited academic / a cited article. We constructed a rich dataset from Portuguese affiliated economists and use zero inflated negative binomial model. This procedure is appropriate for count outcomes, correcting for overdispersion and excess zeros. We also use a fixed effect poisson model to accomodate authors' unobserved heterogeneity.

We analyze results in detail comparing with existing literature and making some theoretical considerations around.

JEL Codes: I23, A10

* NOVA School of Business and Economics

I. INTRODUCTION & LITERATURE REVIEW

Why are some scholars more popular than others in the economics' academia? Probably because they are more cited; or probably they are more cited because they are popular. Independently of the direction of the proposition, it is important to justify why is it relevant to study citations since *this Work Project's goal is precisely to find what is under the veil of being a popular academic, or under a popular article*. For sure citations are not merely an indicator of *popularity* in its most superfluous sense, rather it might cover very important aspects of the academic world. In any case, as Dan Johnson wrote in 1997:

> "All academics want to be cited. There are many reasons for this desire, including the quests for truth, fame or financial rewards. While truth may not be evidenced by citations, the *search* for truth is marked by discussion and deliberation, the very items that citations measure best. Fame in academia is synonymous with citation (...)".

We introduce with the motivation for this paper; we localize and circumscribe our object of study in the *economics of academic research* and revise literature.

Overview of the field: The importance of research is well known but the multiple issues that it raises are also frequently under discussion. Both applied and academic research are unequivocally consider a crucial activity for development, welfare, progress and growth. But at the same time we discuss, on the grounds of the efficient way to foster research, whether there is a wasteful or even an overproduction and trivialization of scholarship in some fields, specially in social sciences. Going this way we find vast economic literature trying to model mechanisms of opportunistic behavior

by scholars, who are not properly committed with scientific advance or looking for the scientific truth, else their motivation is grants, money or reputation oriented. For sure – as in any other job – people tend to react to the latter stimulus, and the role of economics when designing mechanisms of incentives is precisely to harmonize those in order to stimulate – which is different from determine – certain behaviors which are understood to be individually and socially optimal. The relationship between editors, faculties and referees open the field for a discussion on the grounds of moral hazard and political economy of the academe (see Faria (2005), Hamermesh (1994), Frey (2003)). But other topics such as the funding of academe[1] and real effects of academic research, spillovers of knowledge and synergies with applied research (Jaffe, 1989) are usually touched issues on this field. There is also a "more labor economics" subfield literature which intend to explain the determinants of salaries in academe (see for instance Diamond Jr. (1986), Tuckman and Leahey (1975), Hansen, Weisbrod and Straus (1978), with the last two specifically focused in academic economists)[2]. In spite of all these to be important to accommodate our specific question, it is out of our scope to go deep through it. Finally, we find bibliography that shed lights on rankings of institutions, journals and scholars, publishing and citations patterns: bibliometrics.

Bibliometrics: Under this subfield, in which this paper might be included, we can find works about publishing and citations patterns such as the one from Cardoso, Guimarães and Zimmerman (2010) discussing convergence hypothesis in scientific productivity across countries and regions, or the paper by Frey (1992) regarding conceptual differences in the ways faculty conduce their research. Productivity in economics is

1 See for example Mansfield (1995), Salter and Martin (2001) and, for a more sociological approach Benner and Sandstrom (2000).

2 Given the latest studies, it seems to be true the existence of unanimity regarding the significant positive impact of works published in peer reviewed journals, as well as citations, on economists' earnings.

essentially measured by books and peer reviewed articles.[3] Dundar and Lewis (1998) investigated the determinants of productivity distinguishing Biological Sciences, Engineering, Physical Sciences and Mathematics, and Social Sciences. [4]

Still in this subfield, we include explanations of what is behind economists' rankings – or what is under the veil of a large citations' stock for published articles -, which arises several questions and approximates to our point. One that immediately comes up is how to construct those rankings. The variety of reliable ways to build a ranking can have no limits: Kalaitzidakis et al. (2003) is a common reference regarding economic departments' institutions and journals. Coupé (2003) also ranked economists worldwide; the paper provides abundant information about rankings methodology using also past literature on this field. Another common source is the relative impact of economic journals developed by Laband and Piette (1994). But again, to go specifically through rankings methodology is out of the scope of this paper.

What is a citation?

But after all, *what does a citation mean* or *what is a citation?* Indeed, the use of citations to assess individuals, departments and journals is getting even more used as the time comes by (Laband, 1986; Kalaitzidakis et al, 2003; Coupé, 2003) and thus the importance of citations for earnings is becoming also consensual; in the paper we mentioned above, Diamond Jr. aimed to answer "What is a citation worth?", assessing the impact of citations on scholars' earnings, under the light that citations may

3 In other sciences a proxy like number of patents or number of innovations might be used

4 The results show differences across these scientific areas and includes as explanatory variables interesting data such as ratio of graduate students to faculty and differences between public and private institutions. However, it is not possible to see specific coefficients for Economics. Besides this, they also tested the model with citations per average faculty member as dependent variable – instead of journal articles – but besides the already mentioned limitations for our project, this latter model's fit was lower, and significance was also less powerful.

be seen either as a form of recognition for research, or as a proxy for the human capital ability of a researcher, which fosters and proxies the quantity and quality of research output. As the author noted, "If the reward interpretation is correct, then we would expect, holding all else constant, that salaries would be negatively related to citations[5] whereas if the output interpretation is correct we would expect the opposite." By the time of Diamond's paper, evidence showed a positive sign, leading us to think that *one of the interpretations of citations is to consider it as a proxy for output*.

Our aim here is not, however, to discover or model extensively the utility of citations or the channels through which it affects knowledge's advance, despite it is important that before focusing on *what is behind a citation*, we answer *why are they important to study*. Besides form of recognition (non pecuniary reward) and indicator of productivity, and, as a result of the latter, determinacy for earnings, *citations might also be seen as an input*; at a first glance one might be tempted to consider that citations as an input have a negative impact on productivity[6], but it is actually much more plausible that citations follow a self reinforced dynamic (once you are more cited, you indeed have a lower marginal return on next citations, but the effort you have to exert to continue to be cited also may be lower, once now you are more "visible", so in the end the net benefit of might be positive enough to induce you to become more productive). *Signaling quality* might also be a "virtue" of citations: as Johnson noted, "citations may be indicative of notoriety, but simple errors need only one citation to correct them (so will show low value) while complex errors or debates can be considered additions to knowledge in their own right and so are of high value". This also have a kind of

5 This is so because higher salaries and citations are seen as substitute rewards.

6 If there are decreasing marginal returns to citations (Diamond Jr., 1986), whenever there is a high citation value, productivity (understood in this case as articles per period of time) start to fall. Authors would have less incentives to keep the rate of production if the marginal value is decreasing, assuming more articles conduces proportionally to more citations.

efficiency role, since signaling best papers through citations will prevent academics to waste time and other resources looking to unimportant works; off course this gives a strong advantage for those who already have a considerable stock of citations. But it is the way it works: as Laband (1986) noticed, evidence suggests that few economists may exert a dominant influence on advances in economic theory and that often in the academia *what is said is of less importance than who says it*: thus, if you are convinced you have something really important to say to the humanity, it is better to construct a strong *reputation* first. Citations might thus be a good proxy for everything mentioned but obviously they are not a perfect indicator. Consensus of whether article and book's citations might be the best measure of an academic's quality might be strong (Guimarães, 2002; Johnson, 1997; Laband and Piette, 1994), but in any case the issue is discussed and many authors criticize and propose alternatives to measure it because of the possibility of vicious mechanisms (Frey, 2003; Balaban, 1996). We can still mention some articles rejected in the beginning which turned out to be very influential later on (Gans and Shepperd, 1994); the possibility of "endogenous inflation" generated by self citations and / or "salesmanship citations" is also a caveat; the *obliteration phenomena* (too well established science is used and mentioned without being cited anymore) may also bias economists' rankings; finally, and perhaps most important, citations are quite far away from being a measure of the truth embodied in a work[7], but a contribution to stimulate further thought and discussion, which is expected to – paved by the search for the truth - make advances in science.

Given this, to study what is behind citations is a much more relevant work than what could be initially thought. Faculties, departments, Governments and journals

7 In any case we can expect that at the top scientific level, a high degree of at least a *technical truth* is maintained.

have indeed a high interest on this issue for their policy best.

Under the veil of citations: Little is done *explaining what makes at the end of the day an academic economist to be more cited*; most of the papers (already mentioned above) treat citations as exogenous. Quandt (1976) and other earlier authors did some essays, but the data was not abundant. Johnson (1997) investigated the impact of important characteristics such as reputation, documented research, self citations, length, journal quality and experience to explain weighted citations, using a sample with Yale faculties only. The use of standard OLS regression, however, showed a poor econometric performance. A similar methodology was applied in a previous paper that makes a strong background for Johnson (Laband and Sophocleus, 1985). These authors – and discounting for the age of the paper, which relies on a sample of articles published between 1974 and 1976 – found positive and significant, for the purpose of explaining citations, variables such as reputation, if it is a review article, paper length, journal quality and co-authorship. Johnson corrected some of the problems pointed to the latest work such as not considering more than two authors in team produced papers and only consider reputation of first author.

The present Work Project, using data from Portuguese or Portugal's affiliated scholars since 1970, tries not only to provide evidence with new - and larger time interval - data regarding old referenced variables, but also include jointly for the very first time variables such as undergrad and PhD institution, PhD year, field of publishing, title's size, if publish before PhD, some proxies for reputation and co-authorship, number of articles in top journals. Contrary to some of the existing empirics

on this kind of work, we rely on a careful econometric methodology, using a zero inflated negative binomial model and have a clear concern with common caveats in this type of analysis such as unobserved heterogeneity, excess zeros and overdispersion in data. We ran both per author and per article regressions, which allow us to capture different things.

For comparison reasons more details on the existing literature will be mentioned at the time we analyze the results for each variable considered.

Section II and III describe Data, Methodology and Econometric Procedures, respectively. We analyze results in section IV. V concludes.

II. DATA

Related literature commonly stress that available data on this topic is not easily usable, which limits researchers' aims, but indeed it is honest to point out that our final data set is much more rich than most of the mentioned literature's data. But to take out all the potential implicit in data provided was not an easy work.

Our "raw database" was kindly provided by Paulo Guimarães (U. of Porto), who maintains jointly with Miguel Portela (Minho U.) a web site supported by two research centers in Portugal (cef.up and NIPE), where data can be seen and rankings for authors and institutions can be made according to multiple criteria[8]. The database contains detailed information about Portuguese or Portugal's (regardless or affiliation) academic economists, which is crucial for the nature of our work. We have initially listed 1989 articles written by 1483 authors; many of these authors are foreign co-

8 The web site is http://www3.eeg.uminho.pt/economia/nipe/cef.up%2Bnipe-rank/

authors of the Portuguese and thus are not part of the story and there is no information about them[9]; we also have the year and journal of publication, and the number of ISI Web of Knowledge citations' received by each. Journals' articles are selected, as a general rule, from our author's sample that published in journals that belong to the Econlit database and show up in at least two rankings of journals in Economics. This sums 445 journals. There are articles from 1970 to 2010. Each article has associated one or more authors, whose individual information exists for full name, present affiliation, PhD year of completion, PhD and Undergraduate institution and field, date of birth[10]. Information on JEL codes also provided to test for the impact of different subfields.

All the information described was not in one single file: arrive to the final version of the data set required a patient and time consuming merge and programming work[11]. The final file accounts for 1842 articles and 649 authors. A final file with *per article* information naturally resulted in repeated authors and articles IDs, since the same author commonly wrote more than one article and many articles were written by more than one author. For regressions with citations per article as dependent variable this is useful to better capture authors effects. With total authors citations as dependent variable, applying the "collapse" Stata option puts the file with all information *per author*[12]. As mentioned, a considerable number of variables was generated by

9 A big part of this co-authorship is done with non portuguese affiliated scholars, which justifies the huge difference with the final number of authors (649). This also allow us to confirm past evidence that most of co-authorship is doing out of one's own department.

10 There were around 283 observations who were missing this information. In order to avoid the discharge of these we estimated it based on undergrad year, taking off 23 years, which we considered the average age to finish undergrad. After this, 10 observations were still missing and we apply a similar method but using the available PhD year.

11 Another work project could be presented about the database edition!

12 This simply means that variables that were initially referred to articles (for example citations or number of title's characters of an article), are now rearranged for authors (total number of citations that a given author received for all her articles and average number of title's characters of all articles written by that author).

programming the original information. They can be seen in the regression tables in the following section. All data treatment (generation of new variables, merge and collapse of raw databases), as well as regression analysis was made using Stata (version 11.2)[13].

Graphical description of some used variables may be found in the Appendix.

III. METHODOLOGY AND ECONOMETRIC PROCEDURES[14]

A basic methodological question is whether we use total citations of an author or citations of one single article as dependent variable. Both are used and it allows us - more than being a comparison platform - to see different things. Academics probably are more concerned with lifetime or longer period citations than with ones that one single article can garner, hence total citations per author seems more interesting at the beginning to capture the author's research historical. On another level, to make the analysis *per article* is a useful tool to go deeply on other features such as article's specific variables and to see more specific points of authors' career, allowing to lead with disaggregated information. Since many authors wrote more than one article, we can also see whether citations correspond more to authors' or articles' features. Besides, we can also apply a fixed effect model to capture time-invariant unobserved authors' variables (some idiosyncrasy that authors might have, and are not measurable or known). This unobserved heterogeneity would not result in bias if the unobserved variables were uncorrelated with the measured characteristics. But that is very hard to be convincing: most of our collected data (PhD institution, undergrad institution, past

13 Most of the log files are available under request.

14 References for this section are essentially from Cameron and Trivedi (2009). Details on the procedures applied can be found there or in most of existing recent econometrics' book covering Count Data.

reputation, etc) is very likely to covariate with unobserved variables such as talent or ability. To have these different specifications is a powerful tool for our purposes and provides more robustness and openness to our analysis.

As outlined above one of the problems related with past literature on this issue is a poor econometric treatment. Since we face a count random variable, standard practices such as OLS are not appropriate; for example could result in negative counts. Count outcomes are usually non linear and highly non normal and hence a Poisson model – which is designed for discrete non negative outcomes and is intrinsically heteroskedastic – fits better.[15] Equation 1 gives the Poisson probability density function, where $E(Y) = Var(Y) = \mu$ (equidispersion) and Y is citations.

$$\Pr(Y = y) = \frac{e^{-\mu} u^{y}}{y!} \qquad (1)$$

The expected value of citations is given by (2) $\mu = \exp^{Y_{nj}}$, where[16]

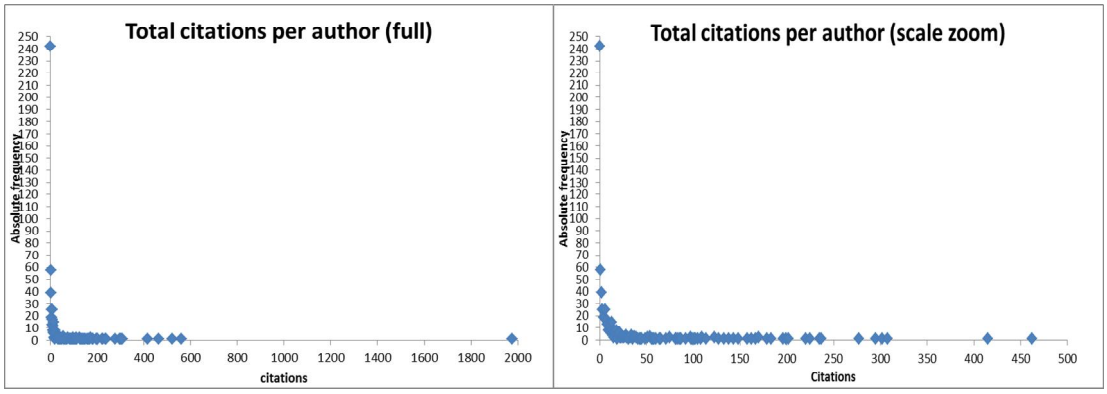
(3) $Y_{nj} = \beta_0 + \beta_1 W_j + \beta_2 X_j + \beta_3 Z_n$ and (4) $\sum_{i=1}^{N} Y_{nj} = \beta_0 + \beta_1 W_j + \beta_2 X_j + \beta_3 \frac{\sum_{i=1}^{N} Z_n}{N} \quad \forall j$.

However, our data reveals a common characteristic in count outcomes that Poisson do not accommodate: overdispersion - the variance is often much larger than the mean. This problem is commonly produced by the presence of unobserved heterogeneity. A more general model – negative binomial – account for overdispersion,

15 We must remember than even if the count is not Poisson distributed, the estimators (MLE) are consistent but require that the conditional mean function $\mu = \exp^{(X'\beta)}$ is correctly specified. The estimation should then be made using pseudo-ML.

16 Y_n stands for article's citations and $\sum_{i=1}^{N} Y_{nj}$ for the lifetime citations of an author. W is a vector of unobservable authors' characteristics, X is a vector of observable authors' characteristics and Z stands for a set of articles' characteristics. There are N articles and J authors; n and j stand for individual article and author, respectively.

but we opted to apply a zero inflated negative binomial (ZINB), since there is a clear excess of zero observations in the variable citations as it can be seen below.



Equation 5 describes the NB (negative binomial) probability mass function NB(u, α)[17].

$$(5)\quad \Pr\left[Y = y \mid \mu, \alpha\right] = \frac{\Psi(\alpha^{-1} + y)}{\Psi(\alpha^{-1})\Psi(y+1)}\left[\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right]^{\alpha^{-1}}\left(\frac{\mu}{\mu + \alpha^{-1}}\right)^{y}$$

As in the case of Poisson $\mu = \exp^{(X'\beta)} \Leftrightarrow \mu = \exp^{Y_{nj}}$, with α constant.

ZINB basically divides the regression in two parts: since the indication of two sub-populations may underlie the counting process, for one of the subpopulations there is a degenerate risk meaning that this individual will observe zero citations, thus we employ a binary model to explain the probability of being zero and a negative binomial for other outcomes. There might be authors who are not cited because they have very "poor" own or articles' characteristics or because independently of their observed characteristics, they will never be cited (similar happens for articles): this introduces a sort of unpredictability in the probability of being cited; we must remember that after all, to be cited is not a self-made process and sometimes it might depend on "idiosyncratic shocks".

17 NB is the closed form of the Poisson-gamma mixture in the special case in which it is the marginal distribution of the sample count, which happens when W – a random variable that causes overdispersion - follows a Gamma(1, α), and α is the variance parameter of the gamma distribution. Note that when α tends to zero, the NB reduces to the Poisson.

Equation 6 provides the density of a zero inflated negative binomial.

$$(6)\quad f(y)=\begin{cases} f_1(0)+\{1-f_1(0)\}f_2(0) & \text{if y=0} \\ \{1-f_1(0)\}f_2(y) & \text{if y}\geq 1 \end{cases}$$

f_1 is the count density of the binary process and the $f_1(0)$ is the probability of being zero for sure. Given the explanation above we will make our binary process only to depend on a constant. The conditional mean is thus:

$$(7)\ \mathrm{E}(y\,|\,\underbrace{W,Z,X}_{A})=\{1-f_1(0\,|\,\text{constant})\}\times\exp(A'\beta)$$

Since in our specifications the binary process – that is, the probability of being zero cited for sure – does not depend on regressors but only on a constant (does not depend on anything observable), then the parameters can be directly interpreted as semi-elasticities, as we do for a common poisson and negative binomial. So, to see marginal effects we know that when a regressor changes one unit, citations change by $\left[e^{\beta}-1\right]\%$, if everything else is kept constant.

Finally, a fixed effect Poisson model is also applied and basically does the same of attributing a dummy for each author – thus authors with only one article are dropped – and so controls possible unobserved forces that can be driving the outcome.

V. RESULTS

Table 1 presents results from the ZINB regression on total authors' citations. The general performance of the estimated model was reasonable: several variables turned to be statistically and economically significant. The p-value of Wald chi-square test (which tests that at least one of the predictors' regression coefficients is not equal to zero) is clearly below 1% which give us a primer good indicator of the model's overall

significance. The model seem to properly accommodate overdispersion, whose existence is evidenced by the dispersion parameter alpha, which is significantly different from zero; this justifies the use of negative binomial (or zero inflated negative binomial) instead of Poisson. I opted to keep ZINB despite tests to diagnose the real need of using it - instead of a standard NB (the Vuong test) – were negative: NB seems to accommodate very well our high number of zeros. In any case, as stated in Cameron and Trivedi (2009), there is no problem whatsoever on using ZINB when NB well accommodates data.

Is it the PhD so important? To hold a PhD is well known as crucial in the academia and econometric analysis is dispensable to know it. But whether a top PhD has strong influence might be more debated. Kocher and Sutter (2001) found that concentration of authors' PhD affiliations on publishing in top journals is substantially higher than the concentration of authors' current affiliations. Results – quite surprisingly – show no significance at 5% on holding a PhD from one of the top 43 universities in the world considered. [18] Evidence from Portugal is thus in contradiction with worldwide: Coupé (2004) report that PhDs of universities with higher reputation have higher probabilities of publishing in economics journals; it would be expected that this higher probability of publishing in top journals also revert in higher citations, but it seems not to be the case and that authors who did not move to a top PhD performed relatively well.[19]

We also controlled for the year of PhD: coefficients' estimates tell us that older PhD graduation increase the probability of being cited. This seems mostly to account for the experience variable and confirm the intuition that more experienced scholars get more citations.

18 Selection criteria is stated in Table 1.

19 Approximately 80% do not hold a top PhD; their average number of citations is 14, compared with 63 from the 20%.

TABLE 1: ZINB Regression Estimates for total citations per author

Total Citations per author			
Regressors	**Aa β**	**P>z**	**ME: dy/dx**
Average year	**-0,0141968**	0,567	-1,41%
#single authored articles	**0,0024676**	0,942	0,25%
#articles with 2/3 authors	**0,1985778**	0,000	21,97%
#articles multiple authored	**0,5055269**	0,000	65,79%
Number Institutions	**0,2687236**	0,003	30,83%
PhD year	**-0,0184439**	0,198	-1,83%
Gender (women=1)	**-0,0810834**	0,577	-7,79%
Nova undergrad=1	**0,3842264**	0,066	46,85%
Católica undergrad=1	**0,7159178**	0,003	104,61%
FEP undergrad=1	**0,0564424**	0,805	5,81%
UTL undergrad=1	**-0,2856754**	0,213	-24,85%
Top PhD=1	**0,2497565**	0,179	28,37%
Publish before PhD =1	**0,0391638**	0,824	3,99%
Reputation (initial period)	**0,3846685**	0,000	46,91%
Reputation (next period)	**-0,0092394**	0,928	-0,92%
_cons	**65,9224**	0,070	
inflate			
_cons	**-17,16059**		
/lnalpha	**0,7276997**		
alpha	2,0703130		
Log pseudo-likelihood	-1.587,692	Obs: 506	
Prob > chi2	0,0000	Zero obs: 163	

Older PhDs (in our dataset it means a degree in the seventies or early eighties) commonly means however that, by those years, to have a PhD was rare: it implied probably no need to publish as much as today to progress in academia, thus it would not be surprising to find a quadratic impact (too old and too recent PhDs would be less cited). It was not modeled in the specification but even if it was probably the impact would not be captured since we control for the number of articles published (whose omission would be the reason for very old PhDs to be less cited). The inclusion of PhD year is possible behind the non significant impact of age, which was so omitted

in the final version. If publish before finish PhD increase citations was also checked: estimates do not have significance; around 45% of our authors published before PhD.

Undergrad institution[20]: It seems not to exist any work explicitly estimating the impact of undergraduate institutions on citations. I identified Nova, Católica, FEP and UTL. Comparing with the impact of having a top PhD may reveal that the first diploma is more important in Portugal to garner citations. Results show the powerful impact of having been studying at Católica and Nova, respectively. Católica reports a marginal impact of more than 100% relatively to other universities not identified; this so strong result is inflated by Rebelo.[21] UTL and FEP are not significant which suggests that the importance of taking a degree in a recognized top national institution *only matters when the jump in prestige or quality is high enough.* Obviously we could refer that we don't know if these effects come indeed from the undergrad institution or if they happen due to self selection, but once the number of observations is relatively large and randomized, the first possibility is more plausible.

Co-authorship: More articles published with one or two colleagues increase average lifetime citations in around 22%. The effect more than doubles if the co-authorship is with more authors but as we can see in the Graph 5, multiple authored articles are very few and so it is dangerous to trust in this conclusion. Several studies have shown an increase in the percentage of papers co-authored over time: we can not know if our estimates in fact correspond to a real co-authorship impact or are due to the fact that who co-authors write more articles. Basically to know the channel that makes co-authorship more cited than publishing alone (if number of articles or higher combination of ideas, etc) is essential. In the end, the co-authorship with 2 or 3

20 The distribution of authors by undergrad institution can be seen in Graph 3 (appendix).

21 Sérgio Rebelo accounts for 1976 citations, while the lifetime average per author sums 22!

coefficient might be inflated by omitted variable bias (number of articles), while the one from multiple authorship might be due to self selection. Anyways, this is the kind of variable it can be very helpful to look at regressions *per article*.

Gender: The paper from Johnson found a significant gender bias, with women facing a higher fixed cost and higher marginal reputation cost. Here, results state that in Portugal women would be quite more cited on average than men but z statistics is in the non rejection region.

Reputation: We build this variable using as proxy the number of published articles in top journals during the first years after the beginning of authors' publishing life.[22] Reputation seems to be one of the main driving forces of total citations per author, increasing it in 46%. Reputation does make a difference. This variable also allow us to see if on average the starting point is important or if there is an initial rejection and whether or not is overcame. Gans and Shepderd (1994) enumerate several very important works that initially were rejected. We enrich our analysis by including a similar variable for a second period, but was reported insignificant: it means the initial jump is the most important[23]. The importance of this variables open avenues for further theoretical work, which could be made around the reputation dynamics of academics. In journals rankings (Kalaitzidakis et al., 2000) the relative positions of top journals ranking constructed using citations from 1998, do not differ that much of a similar study based on 1990 citations. This evidence - in journals – perhaps can be translated to academics: if the starting point and its first evolution is important, then you might start a

22 Ideally we should employ for this variable the number of past citations and subtract them from the regressor; however, this would require also to know the year of each single citation (otherwise we could be accounting with citations of articles that could have been achieved after the initial reputation time have passed) and that information was only available for a very restricted number of articles.

23 Here one could think that this variable may arise endogeneity problems, but that is not the case: indeed publish in top journals depend on past citations, but by the definition we did for initial reputation, we can not contemplate that much significant influence of total citations on number of top publishing papers.

self reinforced process from where it is not easy to be deposed: once better, always better is not is not an improper lemma. However, the second variable related to reputation decrease this impact.

In spite of we have got to control for virtually all variables we hypothesized in theory, off course there are a few more not accounted. One for sure of major importance is whether the faculty accumulates non research related jobs: time constraints can be decisive in the research dedication of a scholar; we did not find any way to control for this, but a broad look at authors in database lead us to intuitively verify that this is true. Another common related variable is teaching. Dundar and Lewis (1998) say that an increase in the teaching load of department is likely to lead to reduced research performance due to time constraints. Becker and Kennedy (2002) contradict this showing that a large fraction of active academic economists do not have difficulties in exemplifying how teaching has for them a far more important role fostering research than existing literature suggests. A few more such as network, department characteristics and funding could be discussed, but some of them may end up to be included in others. For instance, in a lot of cases PhD and undergrad institution can be a proxy of current affiliation, which allow us to control for department characteristics: most of the authors either publish from their graduation department, either from a similar one, or keep in touch with the network created there. The subfield in economics where you publish can also have its influence: in fact we will see right way that it has, when analyzing regression results on citations per article. But in the regression above it would be useless to include number of articles published under each JEL area because we would not be able to identify causality unless we assumed the

crazy assumption that total author's citations are equally distributed per article.

We now present results from the ZINB regression *per article.* As mentioned above, for articles' specific variables we can make a comparison with the poisson fixed effects model: we can see that controlling for author, co-authored articles have no longer a positive significant impact on the number of articles' citations. The latest results evidence that article-invariant unobserved authors' characteristics, that might be correlated with the observed ones, exist and do have some impact.

In this regression we can better capture articles' specific characteristics. We can see that length does matter in a quadratic fashion, but the marginal effect is rather small. Title's size and initial of surname – our joking variables – did not resulted significant, despite whether the surname's initial is "a, b, c or t, u, v, z" reported a relatively large coefficient, despite only significant at 16%. Hollis (2001) have done a similar experiment which also resulted frustrated.

As mentioned there are subfields of economics that might be clearly more likely to be cited than others, either because they are more debated and investigated, or because they are more demanded by journals and then more likely to be cited. Business Economics, Economic Development and Growth are areas that can garner more citations, with Business leading, results we can see to be robust in the fixed effect regression. Economic History and Economic Thought (under the joint code JEL B and JEL N) is the more negative. In any case there is no chance not to admit that these results came clearly biased from the lack of randomization in the portuguese research fields in economics[24]. Johnson also concluded that the higher the number of JEL codes

24 See graph in appendix.

of an article, the lower (and very lower) the number of citations: circumscribed articles are more cited than papers that cover several fields. We did not attempt to verify this.

Table 2: ZINB results with citations per article as dependent variable | Poisson fixed effects

CITATIONS PER ARTICLE				**Poisson Fixed Effects**		
Observations: 2291				Observations 2031		Obs per group
Non zero obs: 1228				Non zero obs 308		Min: 2 Avg: 6,6
Zero obs: 1063				Zero obs 1063		Max: 48
Regressors	**β**	**P>z**	**ME: dy/dx**	**β**	**P>z**	**ME: dy/dx**
Year of publication	-0,145450	0,0000	-13,54%	-0,140198	0,000	-13,08%
Lenght	0,061597	0,0000	6,35%	0,068299	0,000	7,07%
Lenght squared	-0,000540	0,0000	-0,05%	-0,000632	0,016	-0,063%
Size title	-0,005487	0,4480	-0,55%	0,000850	0,881	0,085%
Size title squared	0,000031	0,5150	0,00%	-0,000025	0,464	-0,002%
Top 30 journal=1	0,909049	0,0000	148,20%	0,763283	0,000	114,5%
Co-authorship (2 or 3)=1	0,589084	0,0000	80,23%	0,228216	0,342	25,64%
Multiple authorship (>3)=1	1,087463	0,0000	196,67%	0,197081	0,537	21,78%
PhD year	-0,001060	0,9190	-0,11%			
Gender (women=1)	-0,200403	0,0350	-18,16%			
NOVA undergrad=1	0,391532	0,0020	47,92%			
Catolica undergrad=1	0,236236	0,0470	26,65%			
FEP undergrad=1	0,237903	0,0560	26,86%			
UTL undergrad=1	-0,017135	0,8910	-1,70%			
Top PhD=1	0,223655	0,0170	25,06%			
Publish bef. PhD=1	-0,145532	0,1120	-13,54%			
General Economics	-0,729319	0,2370	-51,78%	-0,444156	0,3380	-35,86%
Maths and Quant. Mehtods	0,064600	0,6190	6,67%	0,366323	0,0350	44,24%
Financial Economics	0,266531	0,0180	30,54%	0,032872	0,7660	3,34%
Law and Economics	-0,000021	1,0000	-0,002%	-0,085253	0,7770	-8,17%
Business	0,905429	0,0000	147,30%	0,844329	0,0000	132,64%
Development, Growth	0,326852	0,0130	38,66%	0,571742	0,0000	77,13%
Economic Systems	-0,117500	0,7250	-11,09%	0,765566	0,0290	115,02%
Micro & IO	0,229166	0,0110	25,76%	0,134969	0,2940	14,45%
Macro, Monetary & Intern. Econ	-0,060317	0,5270	-5,85%	0,241677	0,1300	27,34%
Public & Applied Policy Analys	0,169328	0,0940	18,45%	-0,051190	0,7660	-4,99%
Econ, Hist, & Ec, Thought	-1,065657	0,0010	-65,55%	-1,453752	0,0000	-76,63%
Environmental and Regional Ec.	0,436860	0,0010	54,78%	0,419743	0,0290	52,16%
Misc & Others	0,000487	0,9990	0,05%	-0,470282	0,2260	-37,52%
JEL (miss)	-1,044076	0,0020	-64,80%	-0,916764	0,0040	-60,02%
Reputation (initial)	0,052122	0,0290	5,35%			
Reputation (next period)	0,055634	0,2050	5,72%			
Initial Surname abc=1	0,111951	0,1580	11,85%			
_cons	293,083300	0,0000				
Inflate (logit)						
_cons	-3,207047	0,1280	-0,959524			
/lnalpha	0,980699	0,0000	1,666319			
alpha	2,666319					
Log pseudo-likelihood	-5095,414			-9116,494		
Prob > chi2	0,0000			0,0000		

Another thing that deserves to be mentioned is that the control for top journal publication revealed to be the strongest predictor for citations of an article, a

conclusion that remains above 100% marginal impact in the poisson fixed effects. This specification of regressing articles' citations is the adequate to see the impact of this variable. To do it in the initial one would probably arise endogeneity problems.

Some attention might be given to the coefficient of co-authorship and multiple authorship, dummies relative to single a authored article: in ZINB it largely produces more citations than a single-authored article but in the fixed effects specification both become insignificant: this may indicate that what matters is not if the article is co-authored or not, but with whom is it written.

VI. CONCLUSIONS

The purpose of this work was to explain what is under the veil of being a cited academic, or under a cited article. Citations are indeed important because work as an indicator of academic quality. It is not of less importance to mention the very patient and time consuming work of programming this very rich final data set from the raw data: these files can be very useful for future exploration on this area. Besides, there is no work of this nature that account for so many variables and uses more than very standard econometric procedures. Besides looking for determinants of author's lifetime citations, we also focus on other perspective turning to determinants of articles' cites. A zero inflated negative binomial fitted well to our data and allowed us to correct for overdispersion and excess zeros. A fixed effect poisson provide robustness to results per article accommodating unobserved heterogeneity. Besides not that high, the difference in coefficients reported by the Poisson fixed effect regression is a clear symptom that *very personal characteristics of authors – which we did not observed – do matter to be a cited academic*: to be cited or not to be cited is not just a question of education, time,

field, etc. Talent, ability, propensity to investigate, capacity of produce new ideas and solutions, can be all characteristics that we can not measure or found no data, and are correlated with the observed ones.

Besides, we found that education play an important role garnering citations: specially undergraduate institution revealed to be very important, whenever it has enough quality. To hold a top PhD might be less important than the initial expected, but actually can increase article citations in 25%. Initial reputation also has a crucial mission: it seems that it compensates for those who want to be in the academe to do a strong initial investment signaling their level; then, the process might be self reinforced. We also confirmed the huge importance of publishing in a top journal. We found that to publish in different areas does matter for article citation.

Gender does not affect total citations, but does harm women article's cites.

Co-author seems to gather a lot of citations. The story is more tricky for the effect of publish a single article with other scholar: fixed effects specification seems to indicate that what matters is not if the article is co-authored or not, but with whom is it written.

Despite there is also a component of randomness because of so many variables involved, to be able to explain the coefficients in the way it was done and the robustness of some results also suggests that citations – and thus quality, etc - indeed correspond to something clearly defined. Contrary to some widely spread ideas, it is does not seem to correspond – at least as much as it is passed – to a completed vicious mechanism or a "mafia" clue.

Anyways, as Tuckman and Leahey wrote in 1975 *it should be noted that a distinction exists between the returns to research and the returns to publication.* A

similar distinction might be found for returns to research and returns to citations, whatever those returns are. After all, citations are not everything!

REFERENCES

. Balaban, A. T. 1996. "How should citations to articles in high – and low – impact journals be evaluated, or what is a citation worth?" *Scientometrics*, 37 (3): 495-498

. Becker, W. and Peter Kennedy. 2002. "Does Teaching Enhance Research in Economics?", SSRI working paper

. Cameron, A. Colin, and Pravin K. Trivedi. 2009. "Microeconometrics Using Stata". *Stata Press*

. Coupé, Tom. 2003 "Revealed Performances: World Wide Rankings of Economists and Economic Departments: 1990–2000." Journal of the European Economic Association, 1(6), pp. 1309–1345.

. Coupé, Tom. 2004. "What Do We Know about Ourselves? On the economics of economics." KYKLOS, 57: 197-216

. Dundar, Halil, and Darrel R. Lewis. 1998. "Determinants of Research Productivity in Higher Education." *Research in Higher Education*, 39(6): 607 - 631

. Diamond Jr., Arthur M. 1986. "What is a Citation Worth?" *The Journal of Human Resources*, 21(2): 200-215

. Frey, Bruno S. 1992, "Economics and Economists: an European perspective" AER: Papers and Proceedings, 82(2) 216-220

. Frey, B. S. 2003. "Publishing as Prostitution? - Choosing between one's own ideas and academic success". *Public Choice* 116

. **Faria, João R.** 2005. "The Game Academics Play: Editors vs authors". *Bolletin of Economic Research* 57(1): 307-378

. **Gans and Shepperd** (1994) How are the mighty fallen?

. **Guimarães, Paulo. 2002.** "The state of Portuguese research in economics: an analysis based on publications in international journals." *Portuguese Economic Journal*, 79(5): 957 – 971

. Hammermesh, 1994, Myths and Facts about Refereeing

. **Hollis (2001)**, Co-authorship; *Labor Economics*

. **Hansen, Lee, Burton Weisbrod and Robert Strauss.** 1978. "Modelling the Earnings and Research Productivity of Academic Economists." *Journal of Political Economy*, 86: 729-741

. **Jaffe, Adam B.** 1989. "Real Effects of Academic Research." *American Economic Review*

. Johnson, Dan. 1997. "Getting Noticed in Economics: The Determinants of Academic Citations". *American Economist.* Vol 41, No. 1.

. **Kalaitzidakis et al.** 2003. "Rankings of Academic Journals and Institutions in Economics". *European Economic Review.*

. Laband, D. and M. J. Piette. 1994. "The Relative Impact of Economics Journals: 1970-1990". *Journal of Economic Literature.* Vol. XXXII 640-666.

. Laband, D. and John Sophocleus. 1985. "The Determinants of Article Popularity: Preliminary Results." *Atlanta Economic Journal.*

. Laband, D. 1986. "Article Popularity". Economic Inquiry. Vol. XXIV

. Mansfield (1995), Review of Economics and Statistics

. Salter and Martin (2001) The economic benefits of publicly funded basic research: a critical review Research Policy

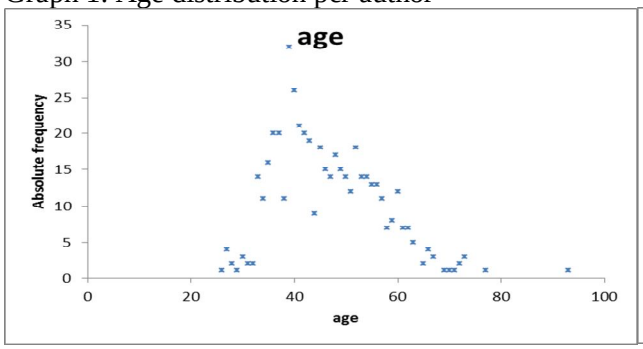
. Tuckman, Howard P., and Jack Leahey. 1975. "What is an article worth?" *The Journal of Political Economy*, 83(5): 951–968.

. Cardoso et al. 2010. "Trends in Economic Research: An international perspective" IZA Discussion Paper

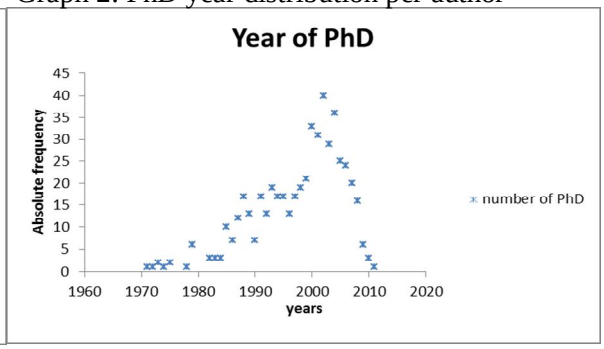
. Quandt, R. E. 1976. "Some Quantitative Aspects of the Journal of Economic Literature". *Journal of Political Economy.* Vol 84(4). p. 741-755

APPENDIX I

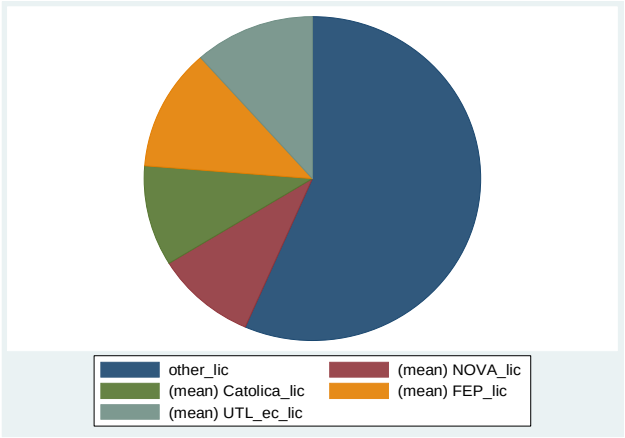
Graph 1: Age distribution per author



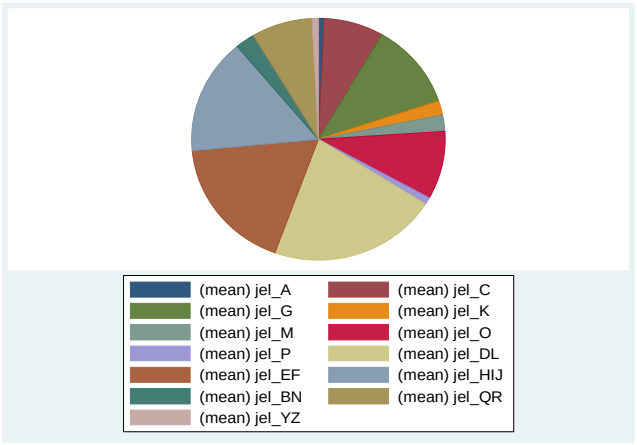
Graph 2: PhD year distribution per author



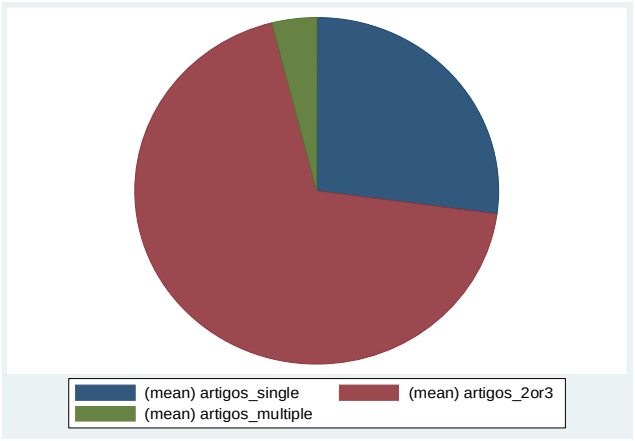
Graph 3: Undergraduate degree distribution per author



Graph 4: share of JEL



Graph 5: Distribution of articles by co-authorship



Artigos_2or3 stand for articles written by two or 3 authors
Artigos_multiple means that articles were written by more than 3 authors

APPENDIX II

Table 4: Authors in database that have not at least a degree directly related with economics

```
. tab autorid if ( (area_lic=="Other Social Sciences" | area_lic=="Engineer") &
> (area=="non Econ") )
```

autorid	Freq.	Percent	Cum.
16	3	18.75	18.75
353	1	6.25	25.00
355	1	6.25	31.25
490	1	6.25	37.50
731	1	6.25	43.75
747	1	6.25	50.00
965	5	31.25	81.25
1116	1	6.25	87.50
1349	2	12.50	100.00
Total	16	100.00	

Table 5: Authors that hold a top PhD

(mean) top_PhD_new	Freq.	Percent	Cum.
0	534	82.28	82.28
1	115	17.72	100.00
Total	649	100.00	

.